CSC 4740/6740 Data Mining

Assignment 4

Due Date: 11:59pm, 11/10/2021

Only the electronic version will be accepted. Please submit it through iCollege.

Suppose David collects the following dataset during 2020.

Table 1. Dataset A.

Index of Records	Rain	Sprinkler	Grass
1	No	No	Dry
2	No	Yes	Wet
3	No	No	Dry
4	No	Yes	Wet
5	Yes	No	Wet
6	No	Yes	Wet
7	No	No	Dry
8	Yes	Yes	Wet
9	No	No	Dry
10	No	Yes	Dry
11	Yes	No	Dry
12	No	No	Dry
13	No	No	Dry
14	No	Yes	Wet
15	Yes	No	Wet
16	No	No	Dry

Suppose David collects another dataset during 2021.

Table 2. Dataset B.

Index of Records	Rain	Sprinkler	Grass
1	No	No	Wet
2	No	No	Dry
3	No	Yes	Dry
4	No	Yes	Wet
5	Yes	Yes	Wet
6	No	Yes	Dry
7	No	No	Dry
8	Yes	No	Wet
9	No	Yes	Wet
10	No	No	Dry

These two datasets A and B are independent. David wants to learn the classification methods by using these two datasets. Suppose attribute “Grass” is the class label in the following classification problems.

Problem 1 (25 points). Suppose David wants to use information gain in the decision tree algorithm. But David does not know how to run decision tree algorithm. Please help him. Please illustrate how to train the decision tree by using the training Dataset A in Table 1.

Solutions:

Problem 2 (25 points). Suppose David wants to use Dataset B in Table 2 as the testing dataset to test the various accuracy values of the decision tree constructed in Problem 1. Please use the decision tree constructed in Problem 1 to test the data records in Table 2. Compare the predicted results with the observed results for “Grass”. Construct the confusion matrix and explain how to compute the following accuracy values.

2.1) Classification accuracy;

2.2) Error rate;

2.3) Sensitivity;

2.4) Specificity;

2.5) Precision;

2.6) Recall;

2.7) F-score;

Solutions:

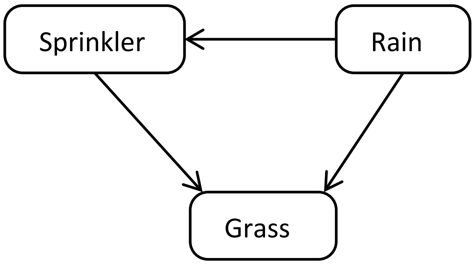
Problem 3 (25 points). Suppose David wants to use Naïve Bayesian Classifier to predict the label (“Grass”) of the fourth record in Dataset B in Table 2:

“Rain” = “No”, “Sprinkler” = “Yes”, and “Grass” = “Wet”.

Please illustrate how to use naïve Bayesian classifier to predict the label (“Grass”) of the data object (“Rain” = “No”, “Sprinkler” = “Yes”).

Solutions:

Problem 4 (25 points). Suppose Susan helps David design the following graphical model.



But Susan forgot to give the conditional probability tables. David does not know how to train this Bayesian Belief network and get those conditional probability tables. Please illustrate how to use Dataset A in Table 1 to train the above graphical model. That is, how to use Dataset A to estimate the probabilities in the conditional probability tables.

Solutions: